arXiv: 1106.6104

DETERMINISTIC SEQUENCING OF EXPLORATION AND EXPLOITATION FOR MULTI-ARMED BANDIT PROBLEMS

BY KEQIN LIU AND QING ZHAO

Department of Electrical and Computer Engineering
University of California, Davis, CA, USA

In the Multi-Armed Bandit (MAB) problem, there is a given set of arms with unknown reward models. At each time, a player selects one arm to play, aiming to maximize the total expected reward over a horizon of length T. An approach based on a Deterministic Sequencing of Exploration and Exploitation (DSEE) is developed for constructing sequential arm selection policies. It is shown that when the moment-generating functions of the arm reward distributions are properly bounded around zero, the optimal logarithmic order of the regret (defined as the total expected reward loss against the ideal case with known reward models) can be achieved by DSEE. The condition on the reward distributions can be gradually relaxed at a cost of a higher (nevertheless, sublinear) regret order: for any positive integer p, $O(T^{1/p})$ regret can be achieved by DSEE when the moments of the reward distributions exist (only) up to the pth order. The proposed DSEE approach complements existing work on MAB by providing corresponding results under a set of relaxed conditions on the reward distributions. Furthermore, with a clearly defined tunable parameter—the cardinality of the exploration sequence, the DSEE approach is easily extendable to variations of MAB, as demonstrated by its generalization to MAB with various objectives and decentralized MAB with multiple players and corrupted reward observations.

1. Introduction.

1.1. *Multi-Armed Bandit.* Multi-armed bandit (MAB) is a class of sequential learning and decision problems with unknown models. In the classic MAB, there are N independent arms and a single player. At each time, the player chooses one arm to play and obtains a random reward drawn i.i.d. over time from an *unknown* distribution. Different arms may have different reward distributions. The design objective is a sequential arm selection policy that maximizes the total expected reward over a long but finite horizon T. The MAB problem finds a wide range of applications including clinical trials, target tracking, dynamic spectrum access,

*Supported by NSF Grant CCF-0830685 and ARO Grant W911NF-08-1-0467.

AMS 2000 subject classifications: Primary 62L05, 93E35; secondary 60G40

Keywords and phrases: multi-armed bandit, deterministic sequencing of exploration and exploitation, logarithmic and sublinear regrets, decentralized multi-armed bandit, corrupted reward observations

Internet advertising and Web search, and social economical networks (see [1, 2, 3] and references therein).

In the MAB problem, each received reward plays two roles: increasing the wealth of the player, and providing one more observation for learning the reward statistics of the arm. The tradeoff between exploration and exploitation is thus clear: which role should be emphasized in arm selection—an arm less explored thus holding potentials for the future or an arm with a good history of rewards? In 1952, Robbins addressed the two-armed bandit problem [1]. He showed that the same maximum average reward achievable under a known model can be obtained by dedicating two arbitrary sublinear sequences for playing each of the two arms. In 1985, Lai and Robbins proposed a finer performance measure, the so-called regret, defined as the expected total reward loss with respect to the ideal scenario of known reward models (under which the arm with the largest reward mean is always played) [4]. Regret not only indicates whether the maximum average reward under known models is achieved, but also measures the convergence rate of the average reward, or the effectiveness of learning. Although all policies with sublinear regret achieve the maximum average reward, the difference in their total expected reward can be arbitrarily large as T increases. The minimization of the regret is thus of great interest. Lai and Robbins showed that the minimum regret has a logarithmic order in T. Furthermore, for Gaussian, Bernoulli, Poisson and Laplacian distributions, policies were explicitly constructed to achieve this minimum regret[1]. Since the seminar work by Lai and Robbins, simpler index-type policies were developed by Agrawal in 1995 [5] and Auer *et al.* in 2002 [6]. These policies achieve the logarithmic regret order under different conditions on the reward distributions.

In the classic policies developed by Lai and Robbins [4], Agrawal [5] and Auer *et al.* [6], arms are prioritized according to two statistics: the sample mean $\bar{\theta}(t)$ calculated from past observations up to time t and the number $\tau(t)$ of times that the arm has been played up to t. The larger $\bar{\theta}(t)$ is or the smaller $\tau(t)$ is, the higher the priority given to this arm in arm selection. The tradeoff between exploration and exploitation is reflected in how these two statistics are combined together for arm selection at each given time t. This is most clearly seen in the UCB1 (Upper Confidence Bound) policy proposed by Auer *et al.* in [6], in which an index $I(t)$ is computed for each arm and the arm with the largest index is chosen. The index (referred to as the upper confidence bound) has the following simple form:

$$I(t) = \bar{\theta}(t) + \sqrt{2\frac{\log t}{\tau(t)}}. \tag{1.1}$$

[1]For the existence of an optimal policy in general, Lai and Robbins established a sufficient condition on the reward distributions. However, the condition is difficult to check and is only verified for the specific distributions mentioned above.

This index form is intuitive in the light of Lai and Robbins's result on the logarithmic order of the minimum regret which indicates that each arm needs to be explored on the order of $\log t$ times. For an arm sampled at a smaller order than $\log t$, its index, dominated by the second term, will be sufficient large for large t to ensure further exploration.

1.2. *Deterministic Sequencing of Exploration and Exploitation.* In this paper, we develop a new approach to the MAB problem. Based on a Deterministic Sequencing of Exploration and Exploitation (DSEE), this approach differs from the classic policies proposed in [4, 5, 6] by separating in time the two objectives of exploration and exploitation. Specifically, time is divided into two interleaving sequences, in one of which, arms are selected for exploration, and in the other, for exploitation. In the exploration sequence, the player plays all arms in a round-robin fashion. In the exploitation sequence, the player plays the arm with the largest sample mean. Under this approach, the tradeoff between exploration and exploitation is reflected in the cardinality of the exploration sequence. It is not difficult to see that the regret order is lower bounded by the cardinality of the exploration sequence since a fixed fraction of the exploration sequence is spent on bad arms. Nevertheless, the exploration sequence needs to be chosen sufficiently dense to ensure effective learning of the best arm. Otherwise, the regret will be dominated by the reward loss in the exploitation sequence caused by incorrectly identified arm rank. The key here is thus finding the minimum cardinality of the exploration sequence that ensures a reward loss in the exploitation sequence having an order no larger than the cardinality of the exploration sequence.

We show that when the moment-generating functions of the reward distributions are properly bounded around zero, DSEE achieves the optimal logarithmic order of the regret using an exploration sequence with $O(\log T)$ cardinality. The condition on the reward distributions can be gradually relaxed at a cost of a higher (nevertheless, still sublinear) regret order. Specifically, we show that for any p, when the moments of the reward distributions only exist up to the pth order, $O(T^{1/p})$ regret can be achieved using an exploration sequence with $O(T^{1/p})$ cardinality. This result reveals an interesting dependency of the regret on the tail probabilities of the reward distributions: a denser exploration sequence is needed when the reward distributions have heavier tails (which makes learning more difficult). In all cases, the regret is sublinear; the maximum average reward defined by the ideal scenario of known reward models is achieved.

Compared to the classic policies proposed in [4, 5, 6] that focus on some specific reward distributions in the exponential family [4, 5] or those with finite support [6], DSEE offers corresponding results under a set of relaxed conditions on the reward distributions. More specifically, the condition on the reward distribu-

tions for achieving the optimal logarithmic regret order is more general than those assumed in [4, 5, 6]. Furthermore, DSEE offers the possibility of sublinear regret for reward distributions with heavy tails. A distinct feature of the DSEE approach is that it has a clearly defined tunable parameter—the cardinality of the exploration sequence—which can be adjusted according to the "hardness" (in terms of learning) of the reward distributions and the observation models. It is thus more easily extendable to handle variations of MAB as discussed in the next subsection.

We point out that for both the classic policies in [4, 5, 6] and the DSEE policies developed in this paper, certain knowledge on the reward distributions is needed for policy construction. In particular, the policies proposed in [4, 5] require the knowledge of the distribution type (*e.g.*, Gaussian or Laplacian) of each arm. The policies proposed in [6] require that the reward distributions have finite support with a known support range. While DSEE achieves the optimal logarithmic regret order for a larger set of reward models, it requires a positive lower bound on the difference in the reward means of the best and the second best arms. This can be a more demanding requirement than the distribution type or the support range of the reward distributions. By increasing the cardinality of the exploration sequence, however, we show that DSEE achieves a regret arbitrarily close to the logarithmic order without *any* knowledge of the reward model. We further emphasize that the sublinear regret for reward distributions with heavy tails is achieved without any knowledge of the reward model (other than a lower bound on the order p of the highest finite moment).

1.3. *Extendability to Variations of MAB.* In the previous subsection, we emphasized the agility of the DSEE approach in handling reward distributions with heavy tails and the lack of any prior knowledge through the adjustment of the cardinality of the exploration sequence. In this subsection, we show that the deterministic separation of exploration from exploitation allows easy extendability of DSEE to decentralized MAB problems.

In a decentralized MAB, there are M $(M < N)$ players. At each time, each player chooses one arm to play. When multiple players choose the same arm, the reward offered by the arm is distributed arbitrarily among the players, not necessarily with conservation. Such an event is referred to as a collision. Players are distributed: actions and rewards of other players are unobservable, and no information can be exchanged among players. As a consequence, collisions are unobservable; a player does not know whether it is involved in a collision, or equivalently, whether the received reward reflects the true state of the arm. Collisions thus not only result in immediate reward loss, but also corrupt the observations that a player relies on for learning the arm rank.

The deterministic separation of exploration and exploitation in DSEE, however,

can ensure that collisions are contained within the exploitation sequence. Learning in the exploration sequence is thus carried out using only reliable observations. Specifically, in the exploration sequence, players play all arms in a round-robin fashion with different offsets which can be predetermined based on, for example, the players' IDs, to eliminate collisions. In the exploitation sequence, each player plays the M arms with the largest sample means calculated using only observations from the exploration sequence under either a prioritized or a fair sharing scheme. While collisions still occur in the exploitation sequence due to the difference in the estimated arm rank across players caused by the randomness of the sample means, their effect on the total reward can be limited through a carefully designed cardinality of the exploration sequence. In particular, we show that under the decentralized policy based on DSEE, the system regret, defined as the total reward loss with respect to the ideal scenario of known reward models and centralized collision-free scheduling among players, grows at the same orders as the regret in the single-player MAB under the same conditions on the reward distributions. These results hinge on the extendability of DSEE to targeting at arms with arbitrary ranks (not necessarily the best arm) and the sufficiency in learning the arm rank solely through the observations from the exploration sequence.

1.4. *Related Work.* The DSEE approach complements the classis results on MAB as detailed in Sec. 1.1 and 1.2. In the context of decentralized MAB with multiple players, the problem was formulated in [7] with a simpler collision model: regardless of the occurrence of collisions, each player always observes the actual reward offered by the selected arm. In this case, collisions affect only the immediate reward but not the learning ability. It was shown that the optimal system regret has the same logarithmic order as in the classic MAB with a single player, and a Time-Division Fair sharing (TDFS) framework for constructing order-optimal decentralized policies using any order-optimal single-player policy as the basic building block was proposed. The same decentralized MAB models as in [7] were also considered in [8, 9] in the context of dynamic spectrum access, where order-optimal distributed policies were established based on UCB1 proposed in [6]. In particular, in [9], UCB1 was extended to targeting at the mth $(1 < m < N)$ best arm by considering both the upper confidence bound given in [6] and a symmetric lower confidence bound. Based on this extension, decentralized polices under both prioritized and fair access scenarios were proposed. In [10], the decentralized MAB with a special imperfect observation model was considered in the context of dynamic spectrum access. Specifically, the arm reward was drawn from the interval $[0, 1]$. Under a collision, either no one gets the reward or the colliding players share the reward evenly. Each player can only observe the received reward. Under a non-cooperative game framework, *i.e.*, each player solely aims to maximize its

own average rewrd. It was shown that the system achieves the maximum long-term average reward when each player adopts the single-player policy proposed in [11]. The system regret order was not considered.

The results presented in this paper and the related work discussed above are developed within the non-Bayesian framework of MAB in which the unknowns in the reward models are treated as deterministic quantities and the design objective is universally (over all possible values of the unknowns) good policies. The other line of development is within the Bayesian framework in which the unknowns are modeled as random variables with known prior distributions and the design objective is policies with good average performance (averaged over the prior distributions of the unknowns). By treating the posterior probabilistic knowledge (updated from the prior distribution using past observations) about the unknowns as the system state, Bellman in 1956 abstracted and generalized the Bayesain MAB to a special class of Markov decision processes [12]. The long-standing Bayesian MAB was solved by Gittins in 1970s where he established the optimality of an index policy—the so-called Gittins index policy [13]. In 1988, Whittle generalized the classic Bayesian MAB to the restless MAB and proposed an index policy based on a Lagrangian relaxation [14]. Weber and Weiss in 1990 showed that Whittle index policy is asymptotically optimal under certain conditions [15, 16]. In the finite regime, the strong performance of Whittle index policy has been demonstrated in numerous examples (see, *e.g.,* [17, 18, 19, 20]).

2. The Classic MAB. Consider an N-arm bandit and a single player. At each time t, the player chooses one arm to play. Playing arm n yields i.i.d. random reward $X_n(t)$ drawn from an unknown distribution $f_n(x)$. Let $\mathcal{F} = (f_1(x), \cdots, f_N(x))$ denote the set of the unknown distributions. We assume that the reward mean $\theta_n \triangleq \mathbb{E}[X_n(t)]$ exists for all $1 \leq n \leq N$.

An arm selection policy π is a function that maps from the player's observation and decision history to the arm to play. Let σ be a permutation of $\{1, \cdots, N\}$ such that $\theta_{\sigma(1)} \geq \theta_{\sigma(2)} \geq \cdots \geq \theta_{\sigma(N)}$. The system performance under policy π is measured by the regret $R_T^\pi(\mathcal{F})$ defined as

$$R_T^\pi(\mathcal{F}) \triangleq T\theta_{\sigma(1)} - \mathbb{E}_\pi[\Sigma_{t=1}^T X_\pi(t)],$$

where $X_\pi(t)$ is the random reward obtained at time t under policy π, and $\mathbb{E}_\pi[\cdot]$ denotes the expectation with respect to policy π. The objective is to minimize the rate at which $R_T^\pi(\mathcal{F})$ grows with T under any distribution set $\mathcal{F}$ by choosing an optimal policy π^*. We say that a policy is order-optimal if it achieves a regret growing at the same order of an optimal policy. We point out that any policy with a sublinear regret order achieves the maximum average reward $\theta_{\sigma(1)}$.

3. The DSEE Approach. In this section, we present the DSEE approach and analyze its performance under different conditions on the reward distributions.

3.1. *The General Structure.* Time is divided into two interleaving sequences: an exploration sequence and an exploitation sequence. In the exploration sequence, the player plays all arms in a round-robin fashion. In the exploitation sequence, the player plays the arm with the largest sample mean calculated from past reward observations. It is also possible to use only the observations obtained in the exploration sequence in computing the sample means. This leads to the same regret order with a significantly lower complexity since the sample means only need to be updated at the same sublinear rate as the exploration sequence. A detailed implementation of DSEE is given in Fig 1 in which only observations from the exploration sequence are used in computing the sample means.

The DSEE Approach

- Notations and Inputs: Let $\mathcal{A}(t)$ denote the set of time indices that belong to the exploration sequence up to (and including) time t. Let $|\mathcal{A}(t)|$ denote the cardinality of $\mathcal{A}(t)$. Let $\overline{\theta}_n(t)$ denote the sample mean of arm n computed from the reward observations at times in $\mathcal{A}(t-1)$. For two positive integers k and l, define $k \oslash l \triangleq ((k-1) \bmod l) + 1$, which is an integer taking values from $1, 2, \cdots, l$.
- At time t,
  1. if $t \in \mathcal{A}(t)$, play arm $n = |\mathcal{A}(t)| \oslash N$;
  2. if $t \notin \mathcal{A}(t)$, play arm $n^* = \arg\max\{\overline{\theta}_n(t),\ 1 \le n \le N\}$.

FIG 1. *The DSEE approach for the classic MAB.*

In DSEE, the tradeoff between exploration and exploitation is balanced by choosing the cardinality of the exploration sequence. To minimize the regret growth rate, the cardinality of the exploration sequence should be set to the minimum that ensures a reward loss in the exploitation sequence having an order no larger than the cardinality of the exploration sequence. This is explicitly stated in the following lemma.

LEMMA 3.1. *Let $R^{\pi}_{T,O}(\mathcal{F})$ and $R^{\pi}_{T,I}(\mathcal{F})$ denote, respectively, the regret incurred in the exploration and exploitation sequences. We have*

$$R^{\pi}_{T}(\mathcal{F}) = R^{\pi}_{T,O}(\mathcal{F}) + R^{\pi}_{T,I}(\mathcal{F}) = \Omega(R^{\pi}_{T,O}(\mathcal{F})). \tag{3.1}$$

There exists an order-optimal policy π^* such that

(i) $R^{\pi^*}_{T}(\mathcal{F}) = O(R^{\pi^*}_{T,O}(\mathcal{F}))$,

(ii) For any policy π with an exploration sequence of cardinality $o(R_{T,O}^{\pi^*}(\mathcal{F}))$, we have $R_T^{\pi}(\mathcal{F}) = \Omega(R_T^{\pi^*}(\mathcal{F}))$.

Furthermore, if (i) and (ii) hold, then π^* is an order-optimal policy.

PROOF. Equation (3.1) is obvious. We consider an order-optimal policy π^* with a specific exploration sequence. If $R_T^{\pi^*}(\mathcal{F}) = \Omega(R_{T,O}^{\pi^*}(\mathcal{F}))$, we can then increase the cardinality of the exploration sequence to the order of $R_T^{\pi^*}(\mathcal{F})$. Since $R_{T,I}^{\pi^*}(\mathcal{F})$ is nonincreasing with the cardinality of the exploration sequence, the regret order remains optimal and (i) holds for this order-optimal policy with a denser exploration sequence. By observing that any exploration sequence with a cardinality order less than the optimal one cannot lead to a better regret, we can see that (ii) also holds. Now we assume that both (i) and (ii) hold for a policy π^*. If there exists a policy achieving a smaller regret order compared to π^*, then (i) contradicts (ii). This leads to the order-optimality of π^*. □

3.2. *The Logarithmic Regret.* In this section, we construct an exploration sequence in DSEE to achieve the optimal logarithmic regret order under the following condition on the reward distributions.

C1. There exist $\zeta > 0$ and $u_0 > 0$ such that $\mathbb{E}[\exp((X-\theta)u)] \leq \exp(\zeta u^2/2)$ for all u with $|u| \leq u_0$.

Conditions C1 implies that the reward distributions have central moments up to an arbitrary order and the diverging rate of the moment sequence is properly bounded. C1 thus imposes constraints on the deviation of the random variable X from its expected value θ, leading to the following Chernoff-Hoeffding bound that states the exponential convergence of the sample mean to the true mean.

LEMMA 3.2. (Chernoff-Hoeffding Bound [21]) *Let $\{X(t)\}_{t=1}^{\infty}$ be i.i.d. random variables drawn from a distribution satisfying C1. Let $\overline{X_s} = (\Sigma_{t=1}^{s} X(t))/s$ and $\theta = \mathbb{E}[X(1)]$. We have, for all $\delta \in [0, \zeta u_0]$, $a \in (0, 1/(2\zeta)]$,*

$$\Pr(|\overline{X_s} - \theta| \geq \delta) \leq 2\exp(-a\delta^2 s). \tag{3.2}$$

Proven in [21], Lemma 3.2 extends the original Chernoff-Hoeffding bound given in [22] that considers only random variables with a finite support. It is not difficult to show that reward distributions in the exponential family (Gaussian, Poisson, Laplacian, and Exponential) as considered in [4, 5, 23] or those with a finite support as considered in [6] satisfy C1. Assuming only C1, DSEE thus offers the optimal logarithmic regret order for a more general set of reward distributions (*e.g.*, the Weibull distribution) as shown in Theorem 3.1 below.

THEOREM 3.1. *Construct an exploration sequence as follows. Let a, ζ, u_0 be the constants such that (3.2) holds. Choose a constant $b > 2/a$, a constant $c \in (0, \theta_{\sigma(1)} - \theta_{\sigma(2)})$, and a constant $w \geq \max\{b/(\zeta u_0)^2, 4b/c^2\}$. For each $t > 1$, if $|\mathcal{A}(t-1)| < N\lceil w \log t \rceil$, then include t in $\mathcal{A}(t)$. Under this exploration sequence, the resulting DSEE policy π^* has regret*

$$R_T^{\pi^*}(\mathcal{F}) \leq C \log T \tag{3.3}$$

for some constant C independent of T.

PROOF. Without loss of generality, we assume that $\{\theta_n\}_{n=1}^N$ are distinct. From the construction of the exploration sequence in π^*, it is easy to see that $R_{T,O}^{\pi^*}(\mathcal{F})$ has a logarithmic order. From (3.1), it suffices to show that $R_{T,I}^{\pi^*}(\mathcal{F})$ has at most a logarithmic order. In particular, based on the Chernoff-Hoeffding bound given in Lemma 3.2, we show that $R_{T,I}^{\pi^*}(\mathcal{F})$ is bounded by some constant independent of T.

During the exploitation sequence, a reward loss happens if the player incorrectly identifies the best arm. To bound $R_{T,I}^{\pi^*}(\mathcal{F})$, we need to bound the number of learning mistakes at the player. Let E_k denote the kth exploitation period which is the kth contiguous segment in the exploitation sequence. We have

$$\begin{aligned} R_{T,I}^{\pi^*}(\mathcal{F}) &= O(\mathbb{E}[\Sigma_{t \notin \mathcal{A}(T)} \mathbb{I}(\pi^*(t) \neq \theta_{\sigma(1)})]) \\ &= O(\Sigma_{k=1}^{\infty} \Pr(\pi^*(t) \neq \theta_{\sigma(1)} \text{ during } E_k)|E_k|). \end{aligned} \tag{3.4}$$

In the following, we bound $\Pr(\pi^*(t) \neq \theta_{\sigma(1)} \text{ during } E_k)$ and $|E_k|$ respectively. Consider $|E_k|$ first. Let $t_k > 1$ denote the starting time of the kth exploitation period. We have

$$|\mathcal{A}(t_k - 1)| = N\lceil w \log t_k \rceil. \tag{3.5}$$

Starting from time t_k, the next exploration period starts at time t if

$$N\lceil w \log(t-1) \rceil \leq |\mathcal{A}(t_k - 1)| < N\lceil w \log t \rceil.$$

Combined with (3.5), we have $t \leq h t_k$ for some constant h. Equivalently, we have

$$|E_k| = t - t_k \leq (h-1) t_k. \tag{3.6}$$

Now we consider $\Pr(\pi^*(t) \neq \theta_{\sigma(1)} \text{ during } E_k)$. We have

$$\Pr(\pi^*(t) \neq \theta_{\sigma(1)} \text{ during } E_k) \leq \Pr(\exists\, 1 < j \leq N \text{ s.t. } \overline{\theta}_{\sigma(j)}(t_k) \geq \overline{\theta}_{\sigma(1)}(t_k)). \tag{3.7}$$

Let $\tau_n(t)$ denote the number of times that arm n has been played during the exploration sequence up to time t. Recall the parameter b defined in the theorem, define

$$\epsilon_n(t) \stackrel{\Delta}{=} \sqrt{(b \log t)/\tau_n(t)} \geq 0.$$

For each bad arm $\sigma(j)$ $(j > 1)$, we have

$$\epsilon_{\sigma(1)}(t_k) = \epsilon_{\sigma(j)}(t_k) < (\theta_{\sigma(1)} - \theta_{\sigma(j)})/2, \tag{3.8}$$

in which the inequality is due to the fact that $\tau_n(t_k) \geq (4b \log t_k/c^2)$ for any n $(1 \leq n \leq N)$. From (3.8), the $\epsilon_{\sigma(1)}(t_k)$-neighbor of the mean of arm $\sigma(1)$ is non-overlapping with that of arm $\sigma(j)$. To bound the possibility of misidentified rank of arm $\sigma(1)$ and arm $\sigma(j)$, it is sufficient to bound the probability of event $\{|\overline{\theta}_s(t_k) - \theta_s| > \epsilon_s(t_k)\}$ for $s \in \{\sigma(1), \sigma(j)\}$. We further observe that, for $s \in \{\sigma(1), \sigma(j)\}$, $\epsilon_s(t_k) \leq \zeta u_0$ due to the fact that $\tau_s(t_k) \geq b \log t_k/(\zeta u_0)^2$. The Chernoff-Hoeffding bound given in Lemma 3.2 is thus applicable (by choosing $\delta = \epsilon_s(t_k)$) and we have, for $s \in \{\sigma(1), \sigma(j)\}$,

$$\Pr(|\overline{\theta}_s(t_k) - \theta_s| > \epsilon_s(t_k)) \leq 2t_k^{-ab}. \tag{3.9}$$

We can then bound (3.7) as follows.

$$\Pr(\pi^*(t) \neq \theta_{\sigma(1)} \text{ during } E_k) \quad \leq \quad gt_k^{-ab} \tag{3.10}$$

for some constant g.

Based on (3.6) and (3.10), we bound (3.4) as follows.

$$\begin{aligned} R^{\pi^*}_{T,I}(\mathcal{F}) &= O(\Sigma_{k=1}^{\infty} \Pr(\pi^*(t) \neq \theta_{\sigma(1)} \text{ during } E_k)|E_k|) \\ &= O(\Sigma_{k=1}^{\infty} t_k^{-ab} t_k) \\ &= O(\Sigma_{t=1}^{\infty} t^{-ab+1}) \quad \text{(note that } ab > 2) \\ &= O(1). \end{aligned} \tag{3.11}$$

We thus proved Theorem 3.1. □

We point out that the policy depends on certain knowledge about the differentiability of the best arm. Specifically, we need a lower bound (parameter c defined in Theorem 3.1) on the difference in the reward means of the best and the second best arms. We also need to know the bounds on parameters ζ and u_0 such that the Chernoff-Hoeffding bound (3.2) holds. These bounds are required in defining w that specifies the minimum leading constant of the logarithmic cardinality of the exploration sequence necessary for identifying the best arm. However, we show that without any knowledge of the reward models, we can increase the cardinality of the exploration sequence of π^* by an arbitrarily small amount to achieve a regret arbitrarily close to the logarithmic order.

THEOREM 3.2. *Let $g(t)$ be any positive increasing sequence with $g(t) \to \infty$ as $t \to \infty$. Revise policy π^* in Theorem 3.1 as follows: include t $(t > 1)$ in $\mathcal{A}(t)$ if $|\mathcal{A}(t-1)| < N\lceil g(t)\log t\rceil$. Under the revised policy π', we have*

$$R_T^{\pi'}(\mathcal{F}) = O(g(t)\log t).$$

PROOF. The proof is similar to that of Theorem 3.1 up to (3.4). It is not difficult to show that equation (3.6) still holds with a different constant. Let $b(t)$ be any positive increasing sequence with $b(t) = o(g(t))$ and $b(t) \to \infty$ as $t \to \infty$. To show (3.8), we choose

$$\epsilon_s(t_k) \stackrel{\Delta}{=} \sqrt{(b(t)\log t)/\tau_s(t)},\ s \in \{\sigma(1), \sigma(j)\}$$

which, after certain deterministic time, become non-overlapping neighbors of $\theta_{\sigma(1)}$ and $\theta_{\sigma(j)}$ due to the fact that $b(t) = o(g(t))$. Based on the same fact, the Chernoff-Hoeffding bound (3.2) is applicable by choosing $\delta = \epsilon_s(t_k)$ and (3.9) still holds (by replacing b by $b(t)$) after certain deterministic time. The proof is then completed by noticing that the quantity ab (now $ab(t)$) in (3.11) becomes larger than 2 after ceratin deterministic time due to the fact that $b(t) \to \infty$ as $t \to \infty$. □

3.3. *Achieving Sublinear Regret under Relaxed Conditions.* In Sec. 3.2, we adopted a condition (C1) on the tail probabilities of the reward distributions $\mathcal{F}$ to ensure that the Chernoff-Hoeffding bound holds. That condition implies the existence of the moments of the reward distributions at any order. In this subsection, we consider a relaxed condition that only requires the existence of the central moments up to a certain order:

C2 There exists a $p > 1$ such that $\mathbb{E}|X - \theta|^p < \infty$.

Under condition C2, the Chernoff-Hoeffding bound does not hold in general. A weaker bound on the deviation of the sample mean from the true mean was established in [24], as given in the lemma below.

LEMMA 3.3. (One-Sided Bound on Boundary Crossings [24]) *Let $\{X(t)\}_{t=1}^{\infty}$ be i.i.d. random variables satisfying $\mathbb{E}|X(1)|^r < \infty$ for some $1 \le r \le 2$ and $\mathbb{E}(X(1)^+)^p < \infty$ for some $p \ge r$. Let $S_k = \Sigma_{s=1}^k X(s)$ and $\theta = \mathbb{E}[X(1)]$. We have, for all $\alpha r > 1$ and $\epsilon > 0$,*

$$\Sigma_{t=1}^{\infty} t^{p\alpha-2}\Pr(\max_{1\le k\le t}(S_k - k\theta) > \epsilon t^{\alpha}) < \infty. \tag{3.12}$$

Based on Lemma 3.3, we have the following probabilistic bound on the deviation of the sample mean from the true mean.

LEMMA 3.4. *Let $\{X(t)\}_{t=1}^{\infty}$ be i.i.d. random variables drawn from a distribution satisfying C2. Let $\overline{X}_t = (\Sigma_{k=1}^{t} X(k))/t$ and $\theta = \mathbb{E}[X(1)]$. We have, for all $\epsilon > 0$,*

$$\Pr(|\overline{X}_t - \theta| > \epsilon) = o(t^{1-p}). \tag{3.13}$$

PROOF. By choosing $\alpha = 1$ and an r $(1 < r \leq \min\{p, 2\})$ in Lemma 3.3, we have, for all $\epsilon > 0$,

$$\Sigma_{t=1}^{\infty} t^{p-2} \Pr(|\overline{X_t} - \theta| > \epsilon) < \infty. \tag{3.14}$$

The double-sided bound holds since both $E(X(1)^+)^p$ and $E(X(1)^-)^p$ exist under C2. By noticing that the term within the summation of the left-hand side in (3.14) is equal to $o(t^{-1})$, we arrive at (3.13). □

Based on Lemma 3.4, we can choose the cardinality of the exploration sequence under DSEE. In the next theorem, we show that by choosing an exploration sequence with cardinality of $O(T^{1/p})$, the system regret with order $T^{1/p}$ can be achieved.

THEOREM 3.3. *Construct an exploration sequence as follows. Choose a constant $v > 0$. For each $t > 1$, if $|\mathcal{A}(t-1)| < vt^{1/p}$, then include t in $\mathcal{A}(t)$. Under this exploration sequence, the resulting DSEE policy π^p has regret*

$$R_T^{\pi^p}(\mathcal{F}) \leq DT^{1/p} \tag{3.15}$$

for some constant D independent of T.

PROOF. Without loss of generality, we assume that $\{\theta_n\}_{n=1}^{N}$ are distinct. Based on (3.1), it is sufficient to show that $R_{T,I}^{\pi^p}(\mathcal{F}) = o(T^{1/p})$. Choose $\epsilon \in (0, \min\{\theta_{\sigma(i)} - \theta_{\sigma(j)} : 1 \leq i < j \leq N\}/2)$. For a t that belongs to the exploitation sequence, define the following event

$$\mathcal{E}(t) \triangleq \{|\overline{\theta}_n(t) - \theta_n| \leq \epsilon, \ \forall\, 1 \leq n \leq N\}.$$

On event $\mathcal{E}$, the player correctly identifies the best arm, *i.e.*, the regret is zero. The regret incurred in the exploitation sequence is thus at the order of the number of time instances at which event $\mathcal{E}(\cdot)$ does not happen. We thus have

$$\begin{aligned} R_{T,I}^{\pi^p}(\mathcal{F}) &= O(\Sigma_{t \notin \mathcal{A}(T), t \leq T} \Pr(\overline{\mathcal{E}(t)})) \\ &= O(\Sigma_{t \notin \mathcal{A}(T), t \leq T} \Sigma_{n=1}^{N} \Pr(|\theta_n - \overline{\theta}_n(t)| > \epsilon)) \\ &= O(\Sigma_{t \notin \mathcal{A}(T), t \leq T} \Sigma_{n=1}^{N} o(|\mathcal{A}(t)|^{1-p})), \end{aligned} \tag{3.16}$$

where the last equality is due to Lemma 3.4.

By the construction of the exploration sequence, for any $t \notin \mathcal{A}(t)$, we have $|\mathcal{A}(t)| \geq vt^{1/p}$. From (3.16), we have

$$R^{\pi^p}_{T,I}(\mathcal{F}) \quad = \quad O(\Sigma^T_{t=1} o(t^{(1-p)/p})). \tag{3.17}$$

We further note that

$$\Sigma^T_{t=1} o(t^{(1-p)/p}) = o(\int_{t=1}^{T} t^{(1-p)/p} dt) = o(T^{1/p}). \tag{3.18}$$

From (3.17) and (3.18), we have

$$R^{\pi}_{T,I}(\mathcal{F}) = o(T^{1/p}).$$

We thus proved the theorem. □

4. Variations of MAB. In this section, we extend the DSEE approach to variations of MAB including MAB under various objectives and decentralized MAB with corrupted reward observations.

4.1. *MAB under Various Objectives.* Consider a generalized MAB problem in which the desired arm is the mth best arm for an arbitrary m. Such objectives may arise when there are multiple players (see the next subsection) or other constraints/costs in arm selection. The classic policies in [4, 5, 6] cannot be directly extended to handle this new objective. For example, for the UCB1 policy proposed by Auer *et al.* in [6], simply choosing the arm with the mth $(1 < m \leq N)$ largest index cannot guarantee an optimal solution. This can be seen from the index form given in (1.1): when the index of the desired arm is too large to be selected, its index tends to become even larger due to the second term of the index. The rectification proposed in [9] is to combine the upper confidence bound with a symmetric lower confidence bound. Specifically, the arm selection is completed in two steps at each time: the upper confidence bound is first used to filter out arms with a lower rank, the lower confidence bound is then used to filter out arms with a higher rank. It was shown in [9] that under the extended UCB1, the expected time that the player does not play the targeted arm has a logarithmic order.

The DSEE approach, however, can be directly extended to handle this general objective. Under DSEE, all arms, regardless of their ranks, are sufficiently explored by carefully choosing the cardinality of the exploration sequence. As a consequence, this general objective can be achieved by simply choosing the arm with the mth largest sample mean in the exploitation sequence. Specifically, assume that a cost $C_j > 0$ $(j \neq m, 1 \leq j \leq N)$ is incurred when the player plays the jth best arm. Define the regret $R^{\pi}_T(\mathcal{F}, m)$ as the expected total costs over time T under policy π.

THEOREM 4.1. *By choosing the parameter c in Theorem 3.1 to satisfy $0 < c < \min\{\theta_{\sigma(m-1)} - \theta_{\sigma(m)}, \theta_{\sigma(m)} - \theta_{\sigma(m+1)}\}$ and letting the player select the arm with the m-th largest sample mean in the exploitation sequence, Theorem 3.1, Theorem 3.2 and Theorem 3.3 hold for $R_T^\pi(\mathcal{F}, m)$.*

PROOF. The proof is similar to those of previous theorems. The key observation is that after playing all arms sufficient times during the exploration sequence, the probability that the sample mean of each arm deviates from its true mean by an amount larger than its non-overlapping neighbor (see (3.8)) is small enough to ensure a properly bounded regret incurred in the exploitation sequence. □

We now consider an alternative scenario that the player targets at a set of best arms, say the M best arms. We assume that a cost is incurred whenever the player plays an arm not in this set. Similarly, we define the regret $R_T^\pi(\mathcal{F}, M)$ as the expected total costs over time T under policy π.

THEOREM 4.2. *By choosing the parameter c in Theorem 3.1 to satisfy $0 < c < \theta_{\sigma(M)} - \theta_{\sigma(M+1)}$ and letting the player select one of the M arms with the largest sample means in the exploitation sequence, Theorem 3.1, Theorem 3.2 and Theorem 3.3 hold for $R_T^\pi(\mathcal{F}, M)$.*

PROOF. The proof is similar to those of the previous theorems. Compared to Theorem 4.1, the condition on c for applying Theorem 3.1 is more relaxed: we only need to know a lower bound on the mean difference between the M-th best and the $(M+1)$-th best arms. This is due to the fact that we only need to distinguish the M best arms from the rest instead of specifying their rank. □

By selecting arms with different ranks of the sample mean in the exploitation sequence, it is not difficult to see that Theorem 4.1 and Theorem 4.2 can be applied to cases with time-varying objectives. In the next subsection, we use these extensions of DSEE to solve a class of decentralized MAB with imperfect reward observations.

4.2. *Decentralized MAB with Corrupted Reward Observations.*

4.2.1. *Distributed Learning and Its Applications.* Consider M distributed players. At each time t, each player chooses one arm to play. When multiple players choose the same arm (say, arm n) to play at time t, a player (say, player m) involved in this collision obtains a potentially reduced reward $Y_{n,m}(t)$ with $\sum_{m=1}^{M} Y_{n,m}(t) \leq X_n(t)$. The distribution of $Y_{n,m}(t)$ can take any unknown form and has any dependency on n, m and t. Players make decisions solely based on

local reward observations without information exchange. Consequently, a player does not know whether it is involved in a collision, or equivalently, whether the received reward reflects the true state ($X_n(t)$) of the arm.

A local arm selection policy π_m of player m is a function that maps from the player's observation and decision history to the arm to play. A decentralized arm selection policy π is thus given by the concatenation of the local polices of all players:

$$\pi_d \triangleq [\pi_1, \cdots, \pi_M].$$

The system performance under policy π_d is measured by the system regret $R_T^{\pi_d}(\mathcal{F})$ defined as the expected total reward loss up to time T under policy π_d compared to the ideal scenario that the collision among the players is avoided through centralized scheduling and $\mathcal{F}$ is known to all players (thus the M best arms with highest means are played at each time). We have

$$R_T^{\pi_d}(\mathcal{F}) \triangleq T\Sigma_{n=1}^{M}\theta_{\sigma(n)} - \mathbb{E}_{\pi}[\Sigma_{t=1}^{T} Y_{\pi_d}(t)],$$

where $Y_{\pi_d}(t)$ is the total random reward obtained at time t under decentralized policy π_d. Similar to the single-player case, any policy with a sublinear order of regret would achieve the maximum average reward given by the sum of the M highest reward means.

One potential application of the problem is dynamic spectrum access in which M secondary users independently search for spectrum opportunities among N channels. The state/reward—0 (busy) or 1 (idle)—of each channel is modeled as an i.i.d. Bernoulli process with unknown mean. At each time, a secondary transmitter chooses a channel to sense and subsequently transmits in this channel if the channel is sensed to be idle. Sensing is assumed to be imperfect: a false alarm or a miss detection can happen at each time. When multiple users transmit in the same channel, they collide and no one transmits successfully (*i.e.*, no one gets reward). The distribution of the reward received and observed by a user thus depends on the number of players that sensed the same channel, which is unknown and may also be time-varying. If a transmission is successful, the receiver sends an ACK back to its transmitter at the end of the transmission. Each secondary transmitter and its receiver need to use the common reward observations (*i.e.*, ACKs) in decision making to ensure synchronous channel selections. The problem thus falls into the imperfect observation model considered here.

Another potential application is multi-agent systems in which M agents search or collect targets in N locations. When multiple agents choose the same location, they share the reward in an unknown way that may depend on which player comes first or the number of colliding players.

4.2.2. *Decentralized Policies under DSEE.* In order to minimize the system regret, it is crucial that each player extracts reliable information for learning the arm rank. This requires that each player collects sufficient observations that are known to have been obtained without collisions. As shown in Sec. 3, efficient learning can be achieved in DSEE by solely utilizing the observations from the exploration sequence. Based on this property, a decentralized arm selection policy can be constructed as follows. Players play all arms in a round-robin fashion with different offsets in the exploration sequence, in which collisions can be eliminated and reliable learning achieved. In the exploitation sequence, each player plays the M arms with the largest sample means calculated using only local observations from the exploration sequence. Specifically, each player distributes the exploitation time to the estimated M best arms based on either a prioritized sharing scheme or a fair sharing scheme. Note that under a prioritized scheme, each player needs to learn the specific rank of one or multiple of the M best arms and Theorem 4.1 can be applied. While under a fair sharing scheme, a player only needs to learn the set of the M best arms (as addressed in Theorem 4.2) and use the common arm index for fair sharing. An example based on a round-robin fair sharing scheme is illustrated in Fig. 2. If the arm index is not common to all players, the entire ranks of the M best arms need to be learned to achieve fair sharing, *e.g.*, using a round-robin schedule with different offsets for playing the arms ordered by their ranks. We point out that under a fair sharing scheme, each player achieves the same average reward at the same rate.

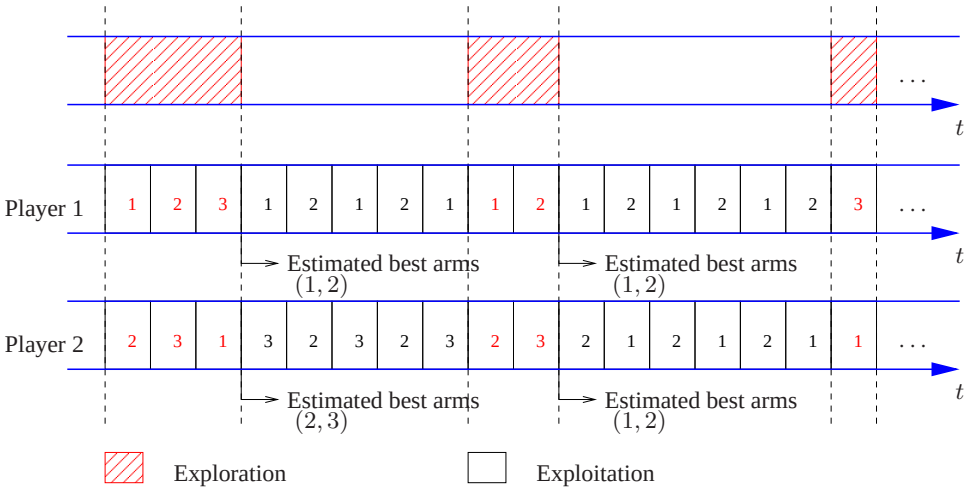

FIG 2. *An example of decentralized policies based on DSEE ($M = 2$, $N = 3$, the index of the selected arm at each time is given).*

THEOREM 4.3. *Under a decentralized policy based on DSEE, Theorem 3.1, Theorem 3.2 and Theorem 3.3 hold for* $R_T^{\pi_d}(\mathcal{F})$.

PROOF. It is not difficult to see that the regret in the decentralized policy is completely determined by the learning efficiency of the M best arms at each player. A detailed proof is thus similar to those of previous theorems. □

5. Conclusion. The DSEE approach addresses the fundamental tradeoff between exploration and exploitation in MAB by separating, in time, the two often conflicting objectives. It has a clearly defined tunable parameter—the cardinality of the exploration sequence—-which can be adjusted to handle reward distributions with heavy tails and the lack of any prior knowledge on the reward models. Furthermore, the deterministic separation of exploration from exploitation allows easy extensions to variations of MAB, including MAB problems under various objectives and with multiple distributed players.

References.

[1] H. Robbins, "Some Aspects of the Sequential Design of Experiments," *Bull. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527-535, 1952.
[2] T. Santner, A. Tamhane, *Design of Experiments: Ranking and Selection*, CRC Press, 1984.
[3] A. Mahajan and D. Teneketzis, "Multi-armed Bandit Problems," *Foundations and Applications of Sensor Management, A. O. Hero III, D. A. Castanon, D. Cochran and K. Kastella, (Editors)*, Springer-Verlag, 2007.
[4] T. Lai and H. Robbins, "Asymptotically Efficient Adaptive Allocation Rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4-22, 1985.
[5] R. Agrawal, "Sample Mean Based Index Policies with $O(\log n)$ Regret for the Multi-armed Bandit Problem," *Advances in Applied Probability*, vol. 27, pp. 1054-1078, 1995.
[6] P. Auer, N. Cesa-Bianchi, P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, vol. 47, pp. 235-256, 2002.
[7] K. Liu and Q. Zhao, "Decentralized Multi-Armed Bandit with Distributed Multiple Players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667-5681, November, 2010.
[8] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed Algorithms for Learning and Cognitive Medium Access with Logarithmic Regret," *IEEE JSAC on Advances in Cognitive Radio Networking and Communications*, vol. 29, no. 4, pp. 781-745, Apr. 2011.
[9] Y. Gai and B. Krishnamachari, "Decentralized Online Learning Algorithms for Opportunistic Spectrum Access," Technical Report, March, 2011. Available at http://anrg.usc.edu/www/publications/papers/DMAB2011.pdf.
[10] G. Kasbekar and A. Proutiere, "Opportunistic Medium Access in Multi-channel Wireless systems: A Learning Approach," in *Proc. of Allerton Conference on Communications, Control, and Computing*, September, 2010.
[11] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The Nonstochastic Multiarmed Bandit Problem," *SIAM J. Computing*, vol. 32, no. 1, pp. 48-77, 2003.
[12] R. Bellman, "A Problem in the Sequential Design of Experiments," *Sankhia*, vol. 16, pp. 221-229, 1956.
[13] J. Gittins, "Bandit Processes and Dynamic Allocation Indices," *Journal of the Royal Statistical Society*, vol. 41, no. 2, pp. 148177, 1979.

[14] P. Whittle, "Restless Bandits: Activity Allocation in a Changing World," *J. Appl. Probab.*, vol. 25, pp. 287-298, 1988.

[15] R. R. Weber and G. Weiss, "On an Index Policy for Restless Bandits," *J. Appl. Probab.*, vol.27, no.3, pp. 637-648, September 1990.

[16] R. R. Weber and G. Weiss, "Addendum to 'On an Index Policy for Restless Bandits," *Adv. Appl. Prob.*, vol. 23, no. 2, pp. 429-430, Jun., 1991.

[17] K. D. Glazebrook, H. M. Mitchell, "An Index Policy for a Stochastic Scheduling Model with Improving/Deteriorating Jobs," *Naval Research Logistics (NRL)*, vol. 49, pp. 706-721, March, 2002.

[18] P. S. Ansell, K. D. Glazebrook, J.E. Niño-Mora, and M. O'Keeffe, "Whittle's Index Policy for a Multi-Class Queueing System with Convex Holding Costs," *Math. Meth. Operat. Res.*, vol. 57, pp. 21-39, 2003.

[19] K. D. Glazebrook, D. Ruiz-Hernandez, and C. Kirkbride, "Some Indexable Families of Restless Bandit Problems," *Advances in Applied Probability*, vol. 38, pp. 643-672, 2006.

[20] K. Liu and Q. Zhao, "Indexability of Restless Bandit Problems and Optimality of Whittle Index for Dynamic Multichannel Access," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547-5567, November, 2010.

[21] R. Agrawal, "The Continuum-Armed Bandit Problem," *SIAM J. Control and Optimization*, vol. 33, no. 6, pp. 1926-1951, November, 1995.

[22] W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13-30, March, 1963.

[23] T. Lai, "Adaptive Treatment Allocation and The Multi-Armed Bandit Problem," *Ann. Statist.*, vol 15, pp. 1091-1114, 1987.

[24] Y. S. Chow, "Delayed Sums and Borel Summability for Independent, Identically Distributed Random Variables," *Bull. Inst. Math. Acad. Sinica*, vol. 1, pp. 207-220, December, 1973.

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
UNIVERSITY OF CALIFORNIA, DAVIS
DAVIS, CA, 95616, USA
E-MAIL: kqliu@ucdvis.edu
qzhao@ucdavis.edu